

# Introduction à la régression

## cours n°1

ENSM.SE – axe MSA

L. Carraro - A. Badea

# Présentation

- **Objectifs**

- Comprendre la problématique générale de la régression.
- Savoir **mettre en œuvre** la régression linéaire dans des cas simples, mais réalistes.
- Comprendre le lien sur un exemple entre analyse de données et régression.
- Ouvrir sur d'autres types de régression (e.g. logistique)

- **Intervenants**

- L. Carraro (LC) [www.emse.fr/~carraro](http://www.emse.fr/~carraro)
- A. Badea (AB)

# Calendrier

- **Mercredi 7 novembre - LC (8h15-11h30)**  
Introduction, retour sur conditionnement, rappels 1A sur un exemple.
- **Mardi 20 novembre - LC (8h15-11h30)**  
ANOVA, interaction et corrélation, résidus et validation.
- **Vendredi 23 novembre - AB (15h15-16h45)**  
TP analyse discriminante.
- **Lundi 26 novembre - LC (8h15-11h30)**  
TP tailles (objectif = savoir-faire de A à Z).
- **Mercredi 28 novembre - AB (8h15-11h30)**  
Retour sur TP analyse discriminante, cours régression logistique.

# Calendrier

- **Mercredi 28 novembre - LC (13h30-15h)**  
Retour sur TP tailles, prévisions, valeurs aberrantes, valeurs influentes.
- **Jeudi 29 novembre - AB (10h-11h30)**  
TP régression logistique.
- **Vendredi 30 novembre - LC (8h15-11h30)**  
TP commun régression/séries temporelles.
- **Lundi 17 décembre (13h30-15h)**  
Examen.

# Logistique

- Documentation :
  - Polycopiés (analyse discriminante, régression linéaire)
  - Photocopies des transparents
- Evaluation
  - TP le 23 novembre : 20%
  - TP le 26 novembre : 30%
  - TP le 29 novembre : 10%
  - Examen avec documents : 40%

# Rappel problématique régression

Lien entre une **réponse**  $y$   
et des **prédicteurs**  $x_1, \dots, x_p$

... dans un contexte incertain

... à partir d'expériences

... dans un but **prédictif**

# Classification

- réponse quantitative/qualitative
- régression paramétrique/non paramétrique
- régression linéaire/non linéaire
- prédicteurs contrôlés/non contrôlés

# Aspects formels

- L'espérance conditionnelle comme espérance
- L'espérance conditionnelle comme projection
- L'espérance conditionnelle linéaire
- Cas gaussien
- Prédicteurs qualitatifs

# Rappel 1A

- Le modèle linéaire et ses hypothèses :
  - indépendance
  - homoscedasticité
  - normalité
- Estimation des paramètres
  - Equation normale (moindres carrés)
  - Propriétés statistiques
- Tests sur les paramètres

# Le modèle linéaire

- $Y$  vecteur des réponses
- $X$  matrice du plan d'expériences
- $\beta$  vecteur des paramètres
- $\varepsilon$  vecteur des écarts au modèle :

$$Y = X\beta + \varepsilon$$

$\varepsilon_1, \dots, \varepsilon_n$  sont des réalisations de v.a.

$E_1, \dots, E_n$  indépendantes et de loi  $N(0, \sigma^2)$

# Estimation et propriétés de $\hat{\beta}$

- Equation normale :  $X'X \hat{\beta} = X'Y$
- $\hat{\beta}$  est de loi normale.
- $E(\hat{\beta}) = \beta$
- $\text{Cov}(\hat{\beta}) = \sigma^2 (X' X)^{-1}$
- $\hat{\beta}$  est BLUE
- A un facteur près,  $\text{cov}(\hat{\beta})$  ne dépend que du plan d'expériences, pas des résultats.

# Tests sur les paramètres

- Hypothèse nulle  $H_0 : \beta_i = 0$
- Statistique de test :

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)}$$

- $T$ , appelée t-ratio, est de loi de Student  $t_{n-p-1}$
- Décision à partir de la p-valeur

# Exemple 1

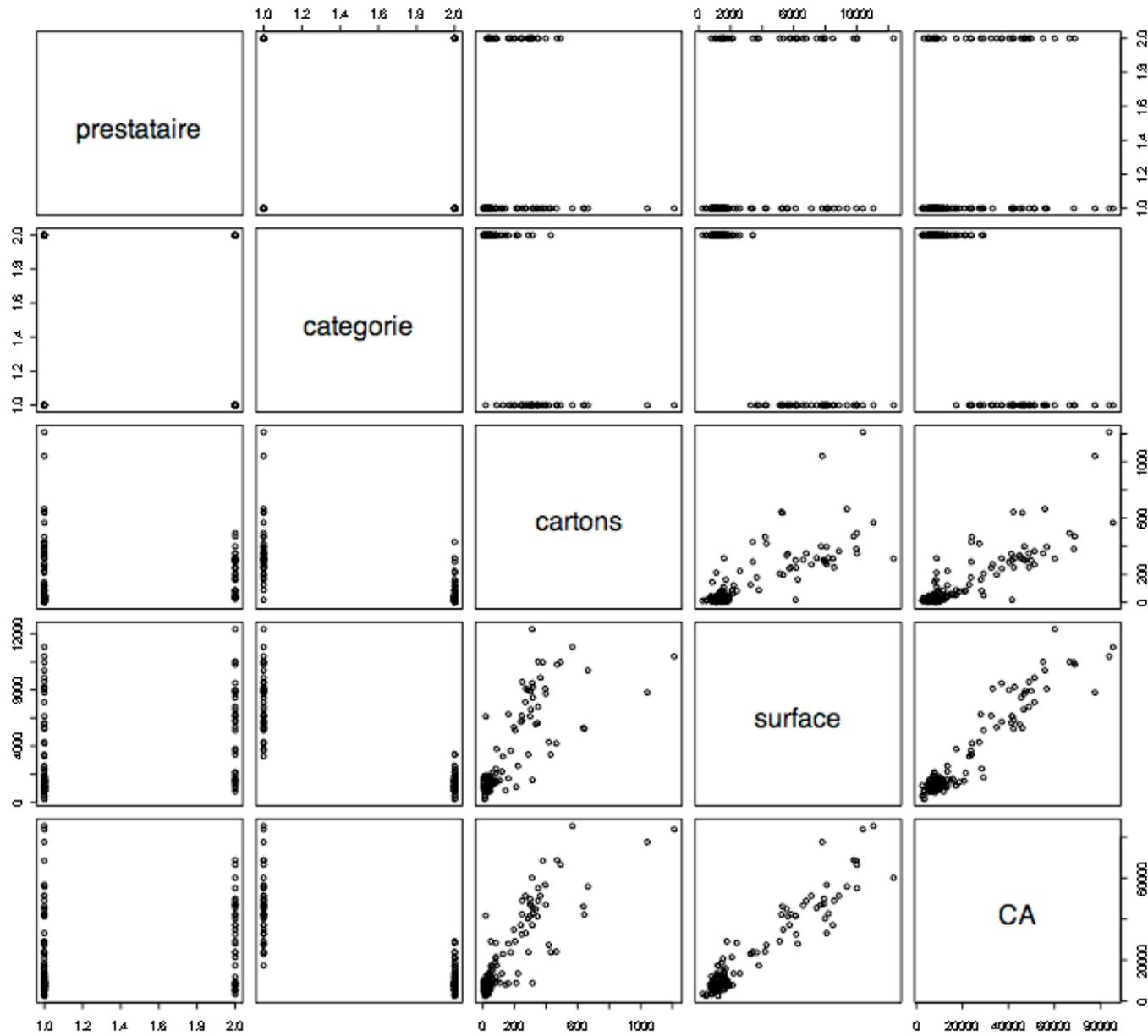
## Récupération de cartons dans la grande distribution

- Réponse : tonnage de cartons annuel
- Prédicteurs :
  - CA HT
  - surface commerciale
  - prestataire (A ou B)
  - catégorie (super ou hyper)
- 137 observations

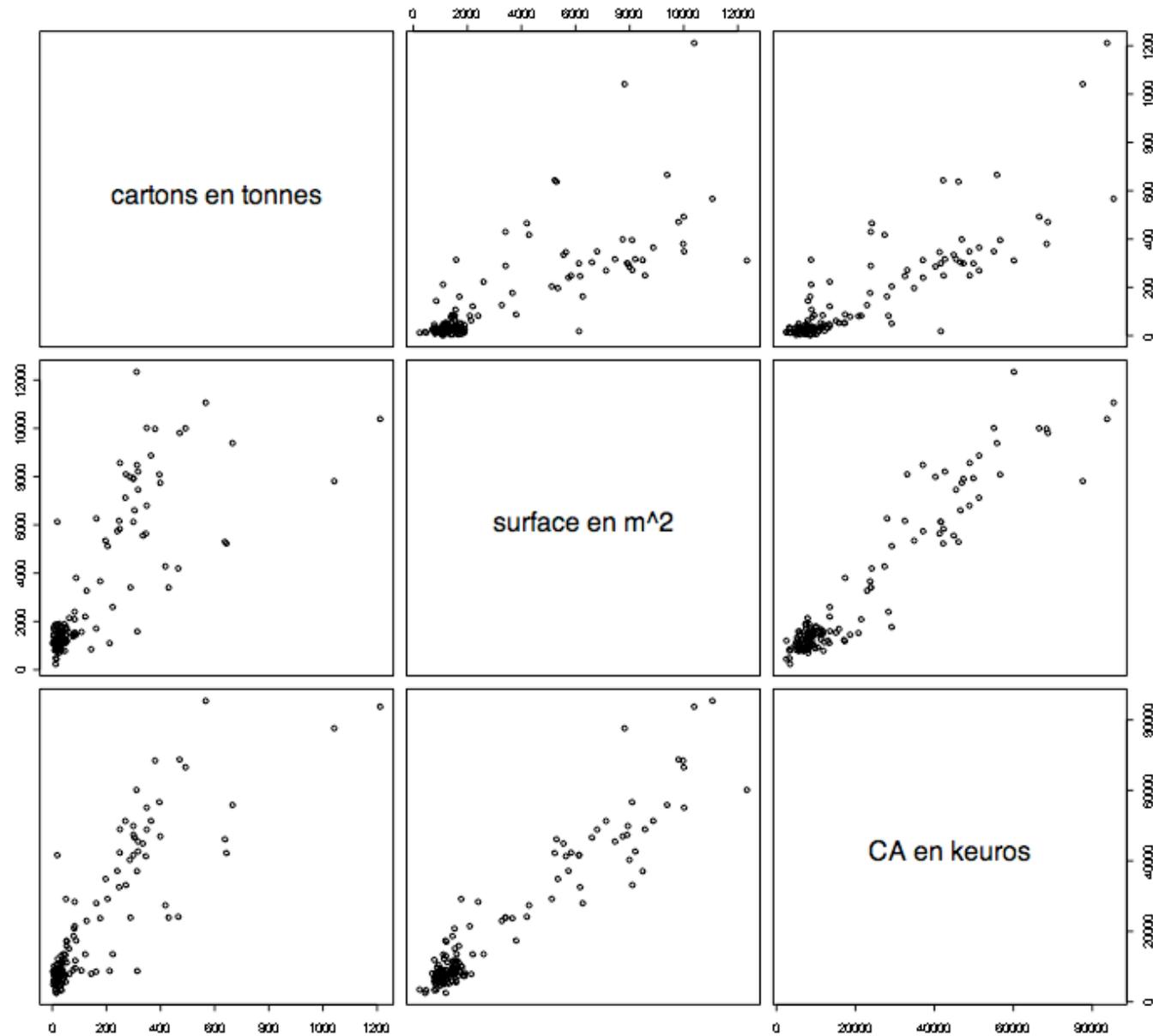
# Questions du client

- Peut-on prévoir le tonnage cartons produit ?
  - Si oui, avec quelle précision ?
- Les deux prestataires ont-ils des résultats analogues ?
  - L'info tonnage vient des prestataires.
  - Approche « toutes choses égales par ailleurs ».

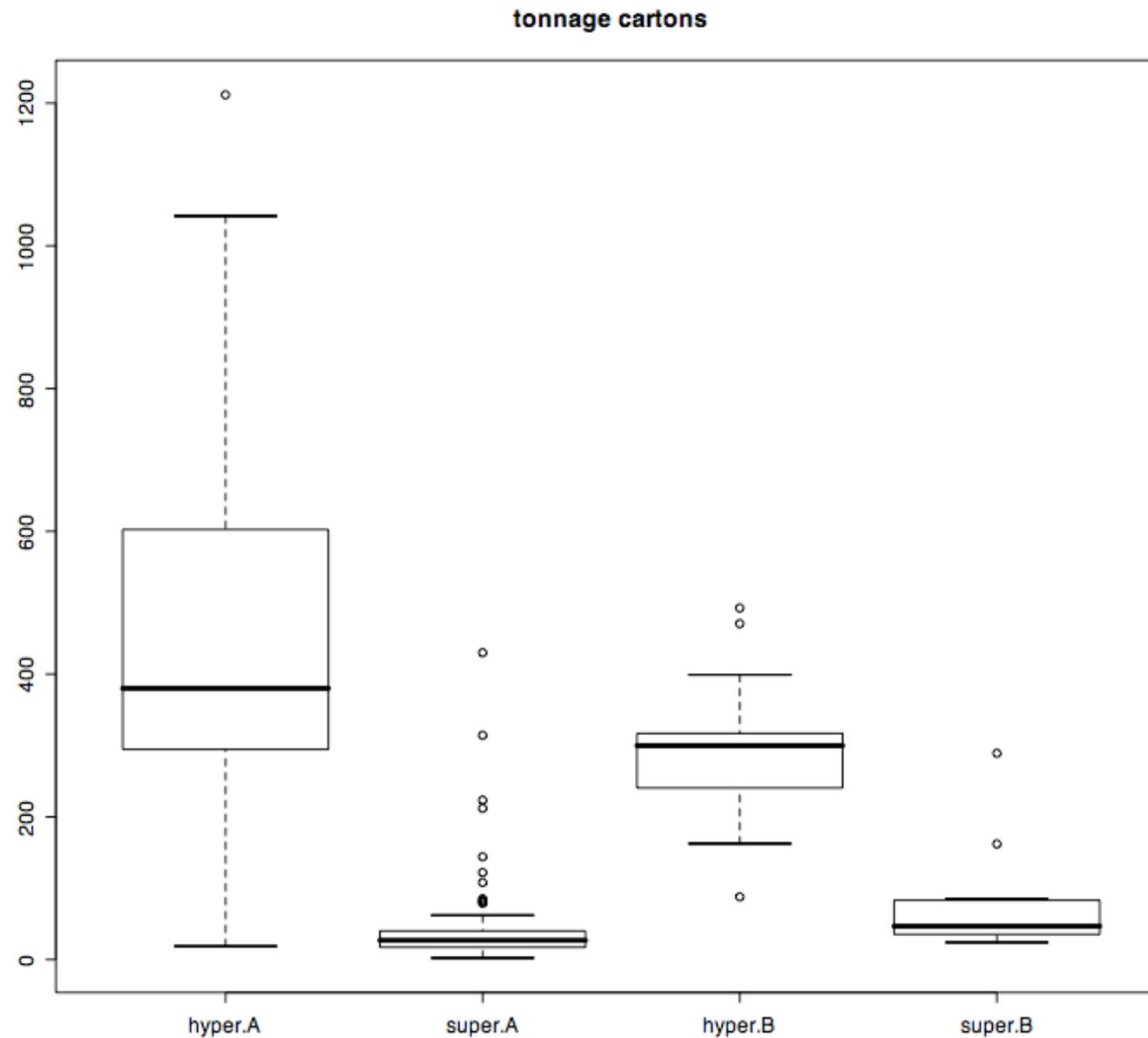
# plot(dechets)



`pairs(cartons~surface+CA, data=dechets)`



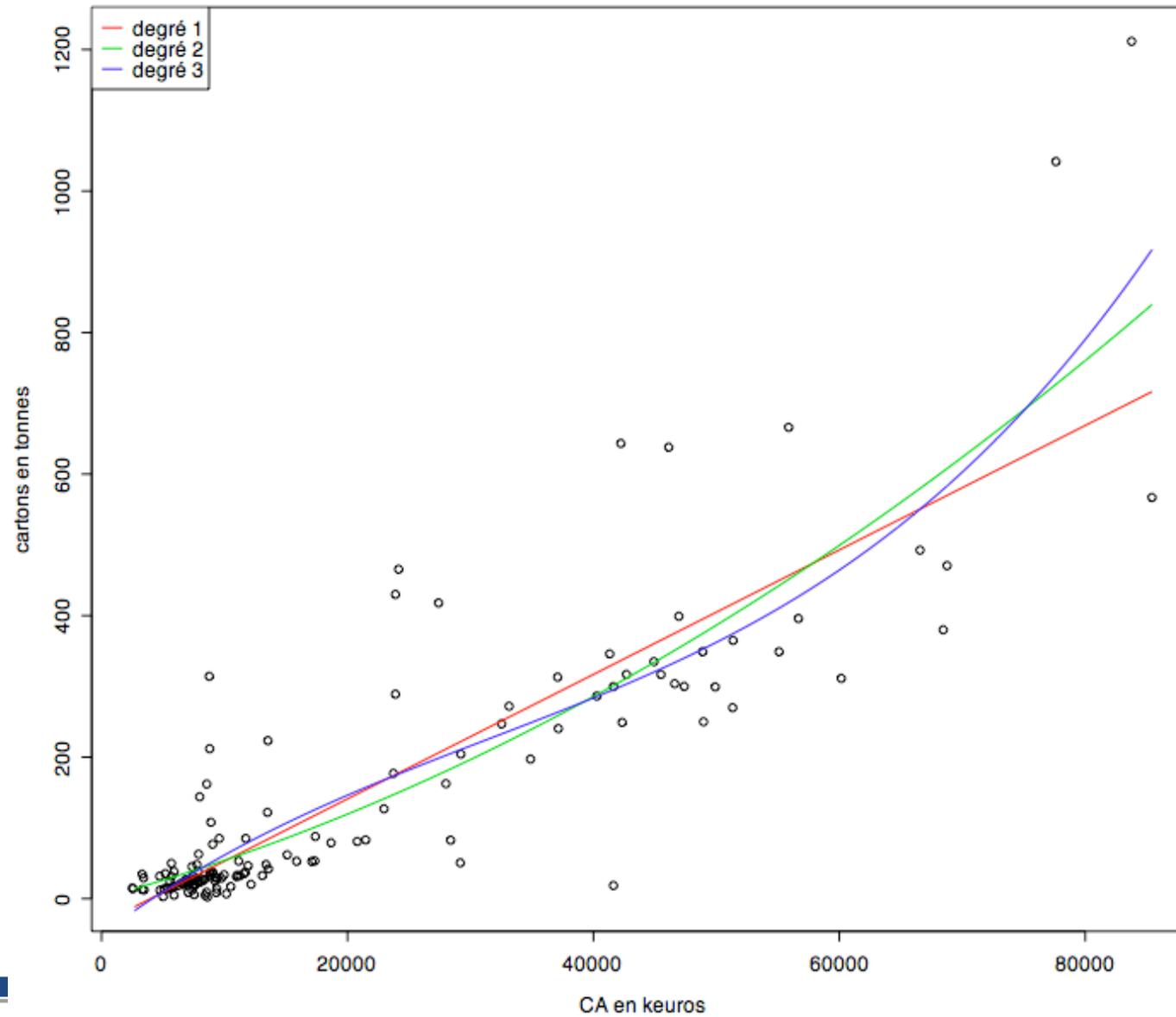
```
boxplot(cartons~categorie+prestataire, data=dechets)
```



# Premières observations

- CA et Surface semblent influentes
- Catégorie et Prestataire aussi
- Questions :
  - Faut-il considérer toutes les variables (on observe une corrélation entre CA et Surface) ?
  - Modélisation de quel type ?

# modèles polynomiaux : cartons $\sim f(\text{CA})$



# Premier modèle

```
lm(formula = cartons ~ CA + surface + categorie + prestataire,
    data = dechets)
```

Residuals:

Min	1Q	Median	3Q	Max
-335.36	-43.78	-12.45	16.03	485.52

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	25.722738	54.069053	0.476	0.6350
CA	0.010893	0.001506	7.235	3.41e-11 ***
surface	-0.020421	0.011541	-1.769	0.0791 .
categoriesuper	-49.420383	45.824569	-1.078	0.2828
prestataireB	-22.096807	22.645039	-0.976	0.3310

---

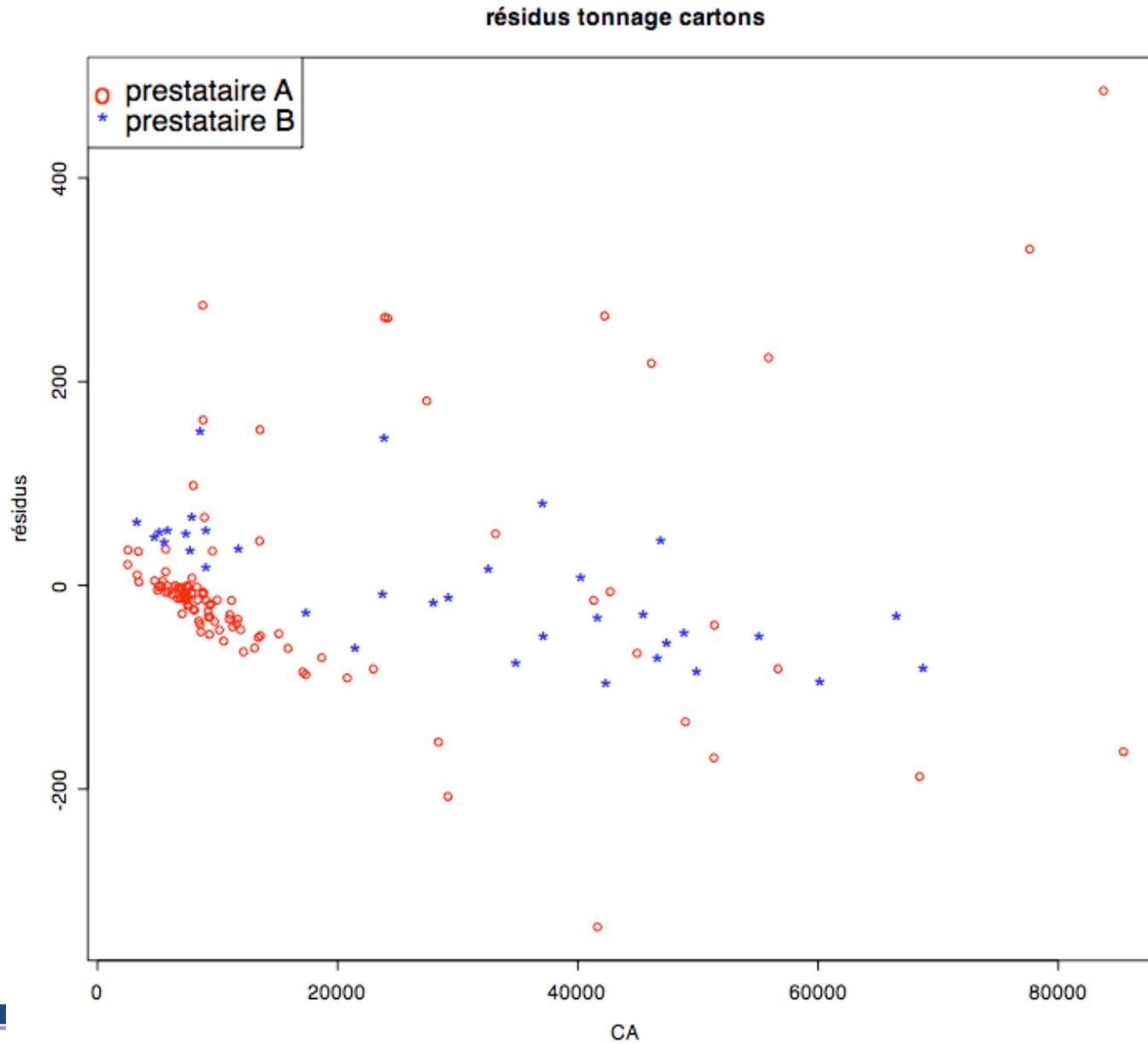
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 101.3 on 132 degrees of freedom

Multiple R-Squared: 0.7451, Adjusted R-squared: 0.7373

F-statistic: 96.45 on 4 and 132 DF, p-value: < 2.2e-16

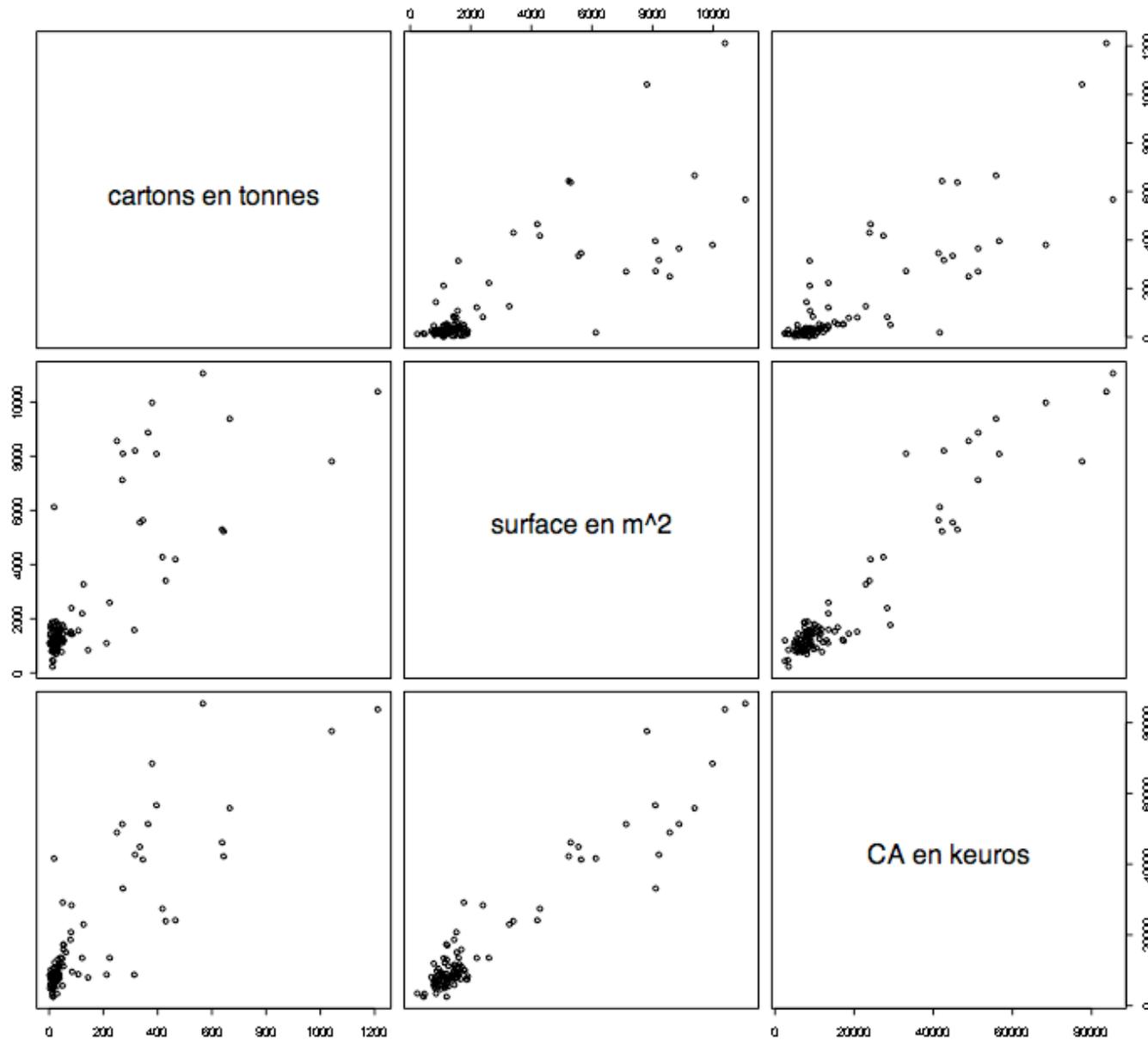
# résidus premier modèle



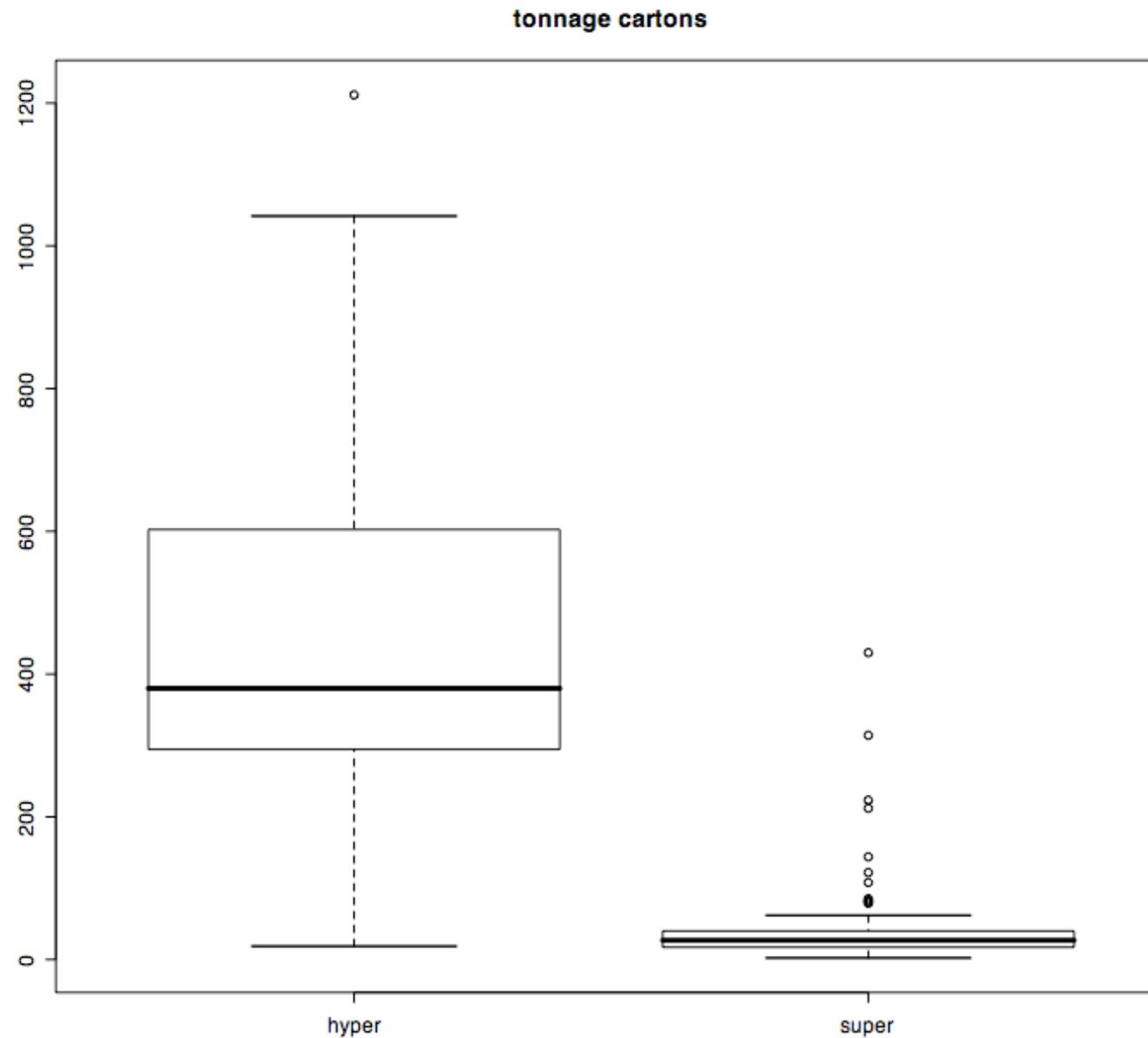
# Remarques

- Résidus de variance dépendant du prestataire
- Problème de biais pour de petits CA
- Deux décisions :
  - Séparer les données par prestataire
  - Passage éventuel au log

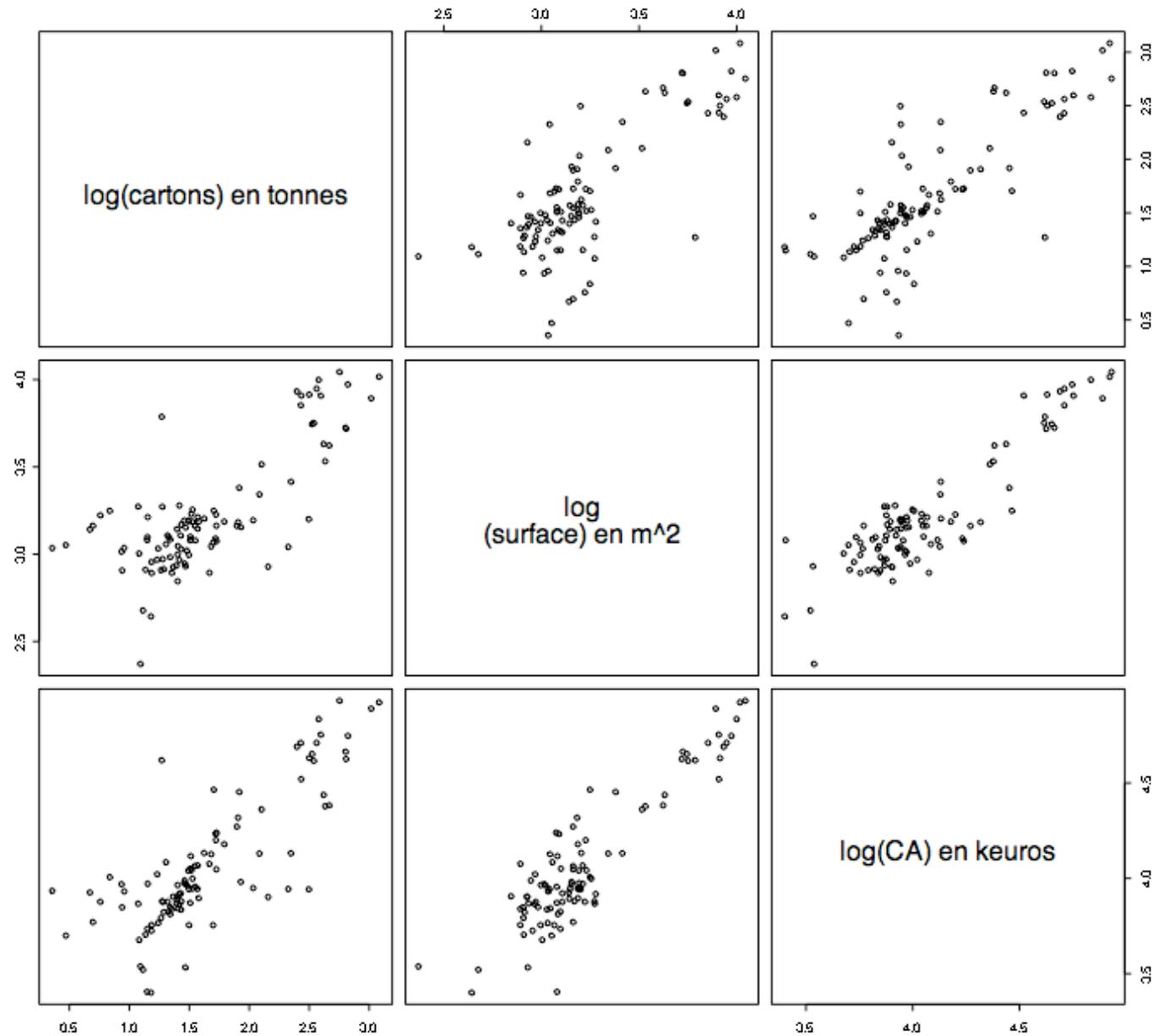
# pairs(cartons~surface+CA) - prestataire A



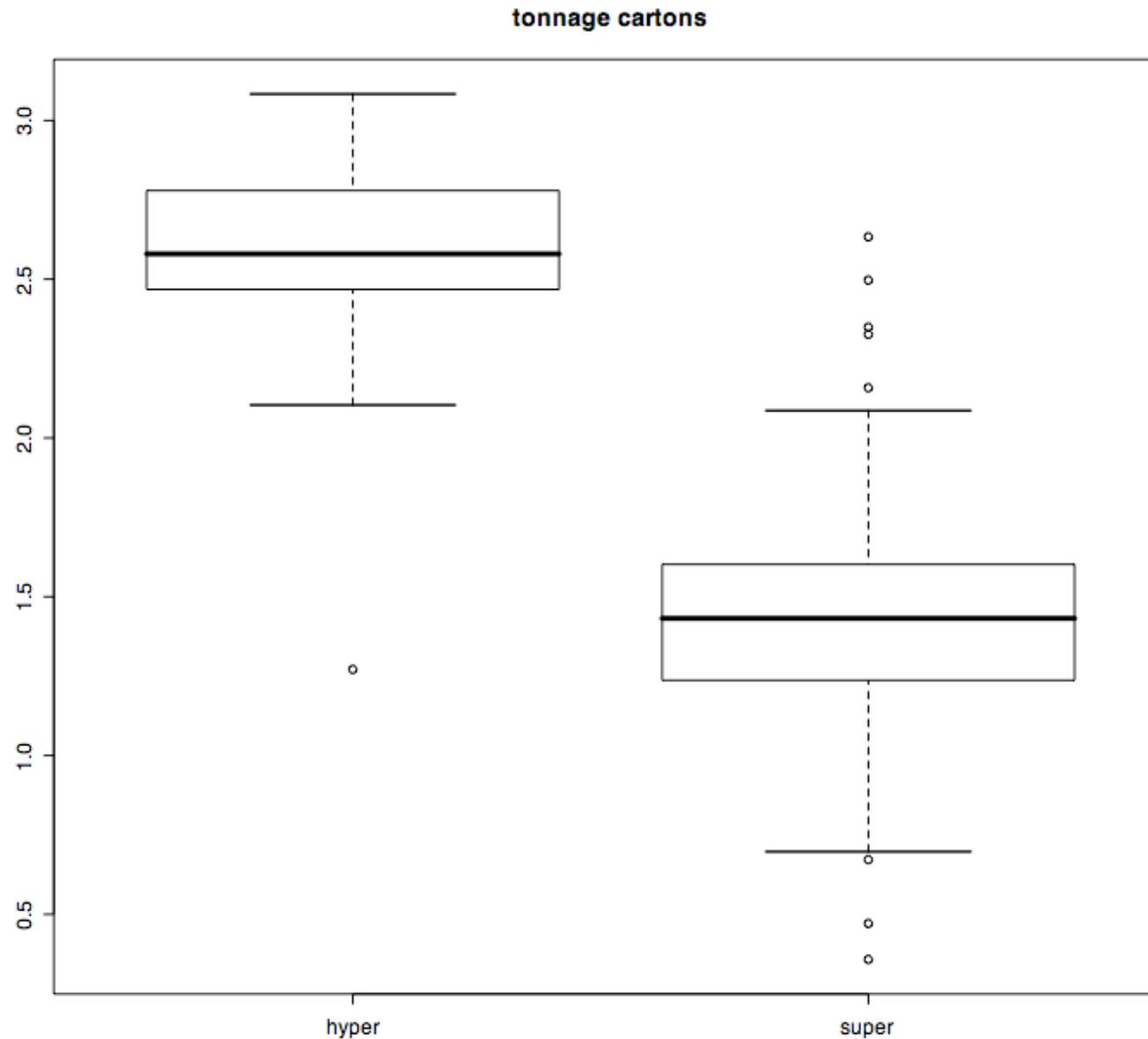
# boxplot(cartons~categorie) - prestataire A



# pairs(cartons~surface+CA) - prestataire A- log10



# boxplot(log10(cartons)~categorie) - prestataire A



# Modèle pour les log - prestataire A

```
lm(formula = log10(cartons) ~ log10(CA) + log10(surface) +
    categorie,
    data = dechetsA)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.22837	-0.07155	-0.01158	0.11457	1.02941

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-2.4924	0.9084	-2.744	0.00722	**
log10(CA)	0.9725	0.2382	4.082	9.1e-05	***
log10(surface)	0.1320	0.2809	0.470	0.63941	
categoriesuper	-0.2964	0.1857	-1.596	0.11372	

---

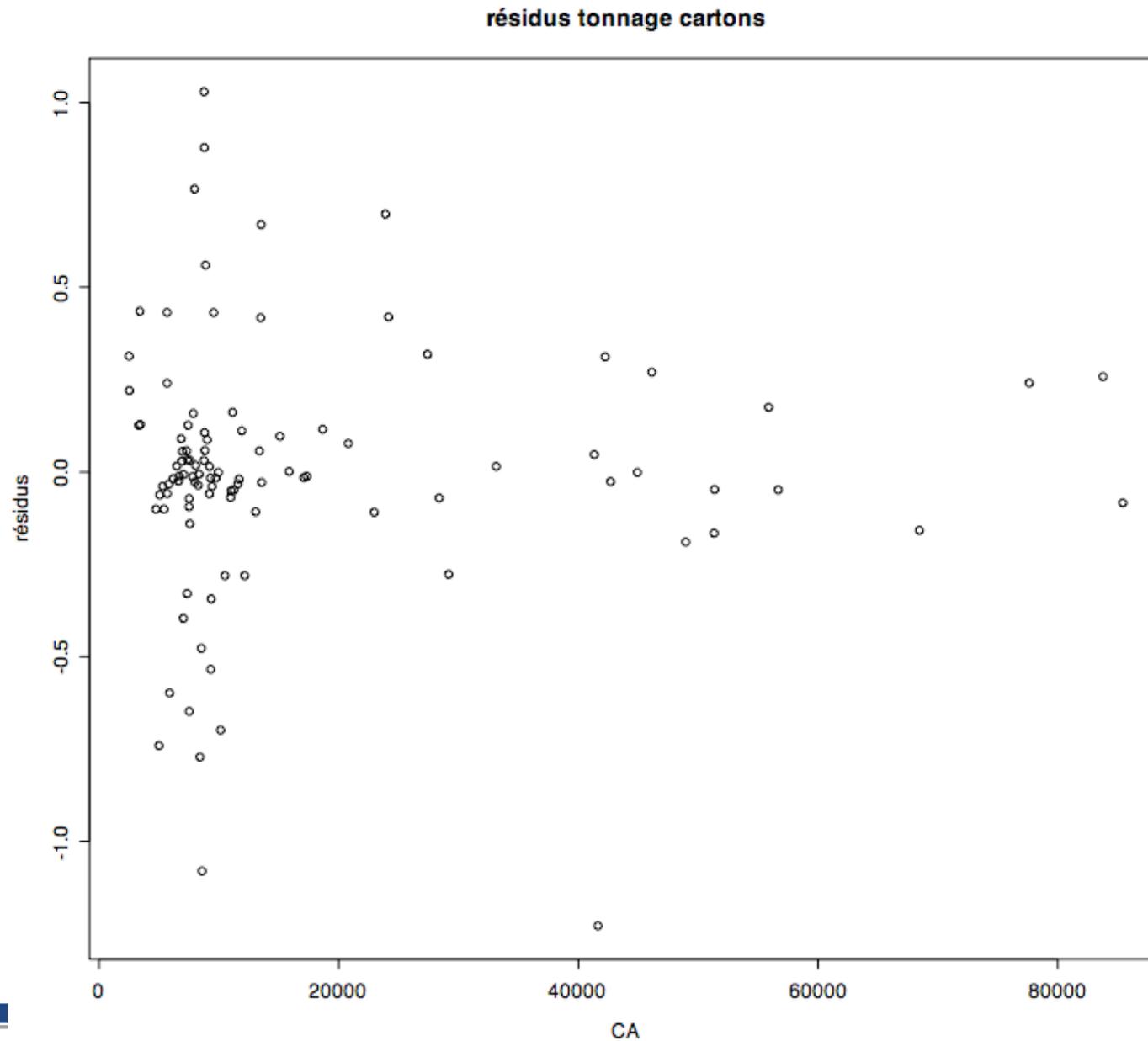
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3474 on 98 degrees of freedom

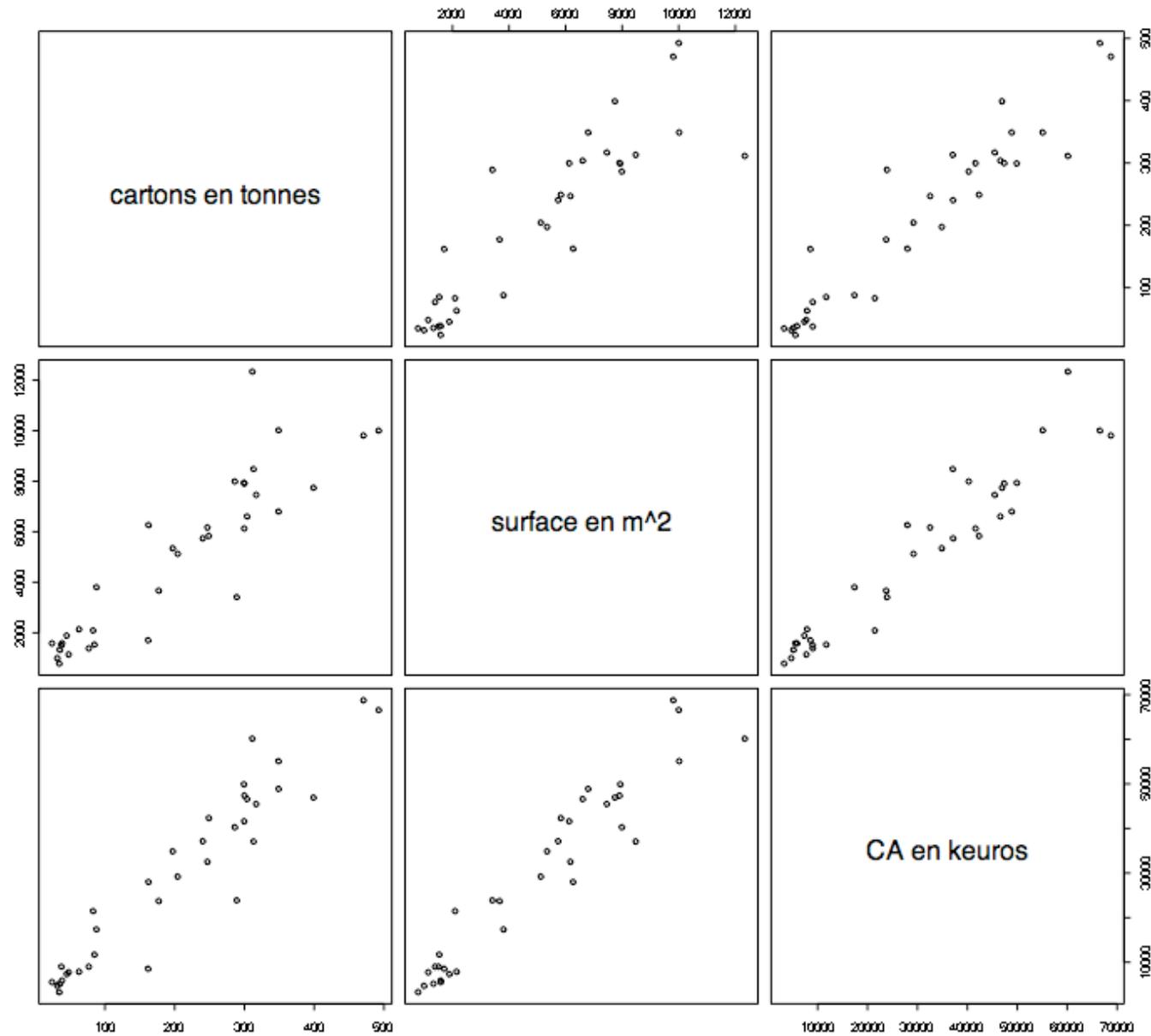
Multiple R-Squared: 0.6625, Adjusted R-squared: 0.6521

F-statistic: 64.12 on 3 and 98 DF, p-value: < 2.2e-16

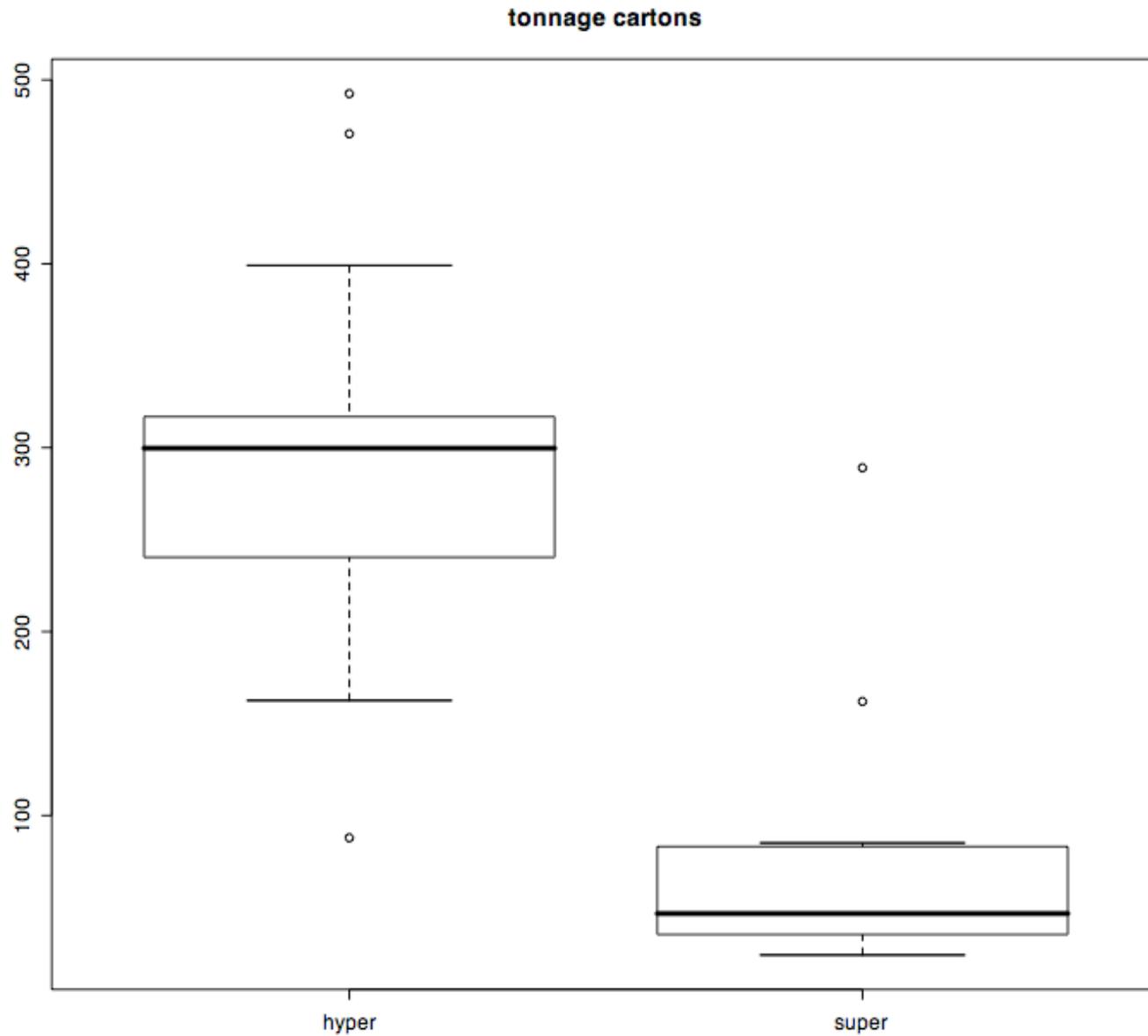
# résidus de ce modèle - prestataire A



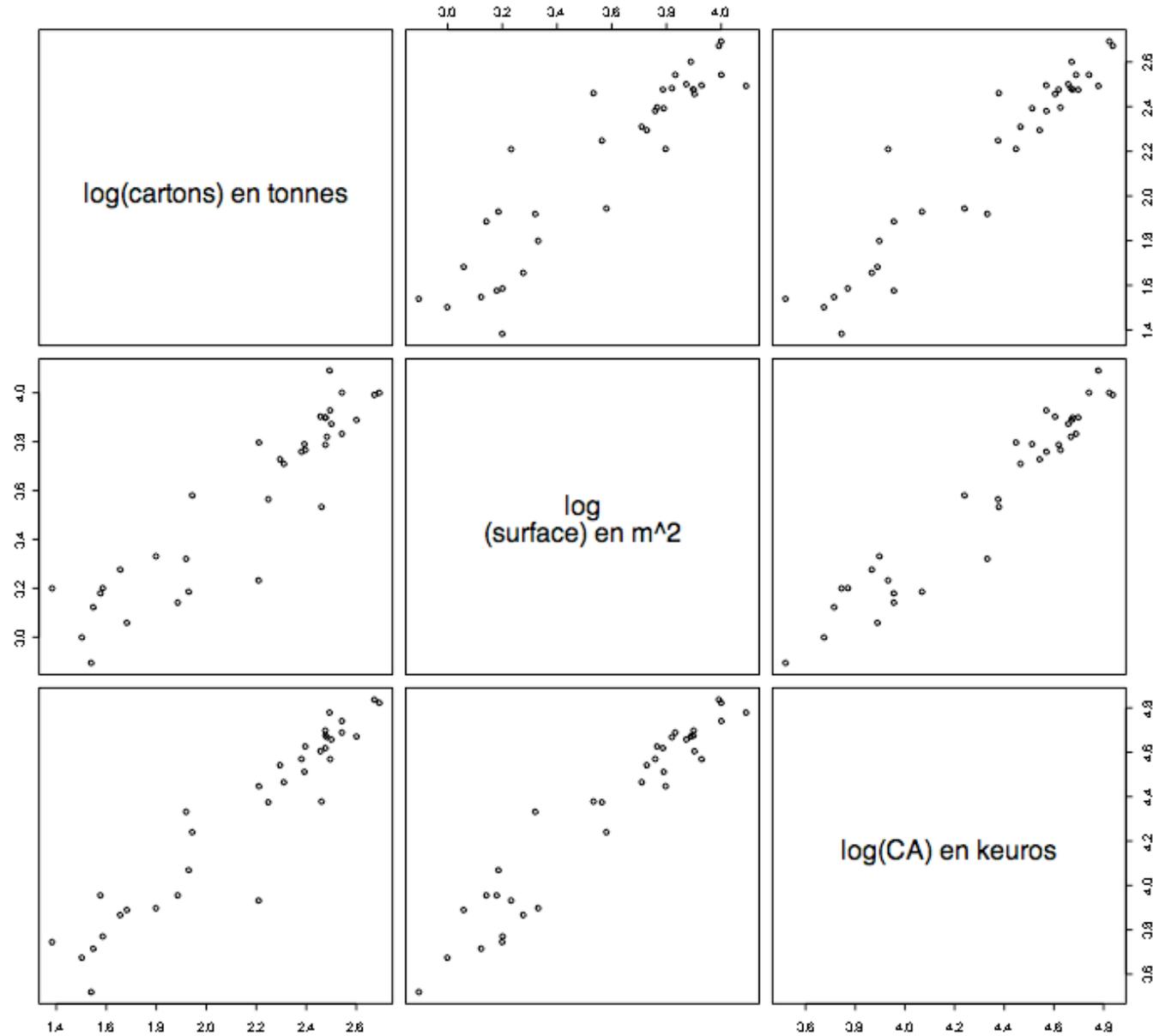
# pairs(cartons~surface+CA) - prestataire B



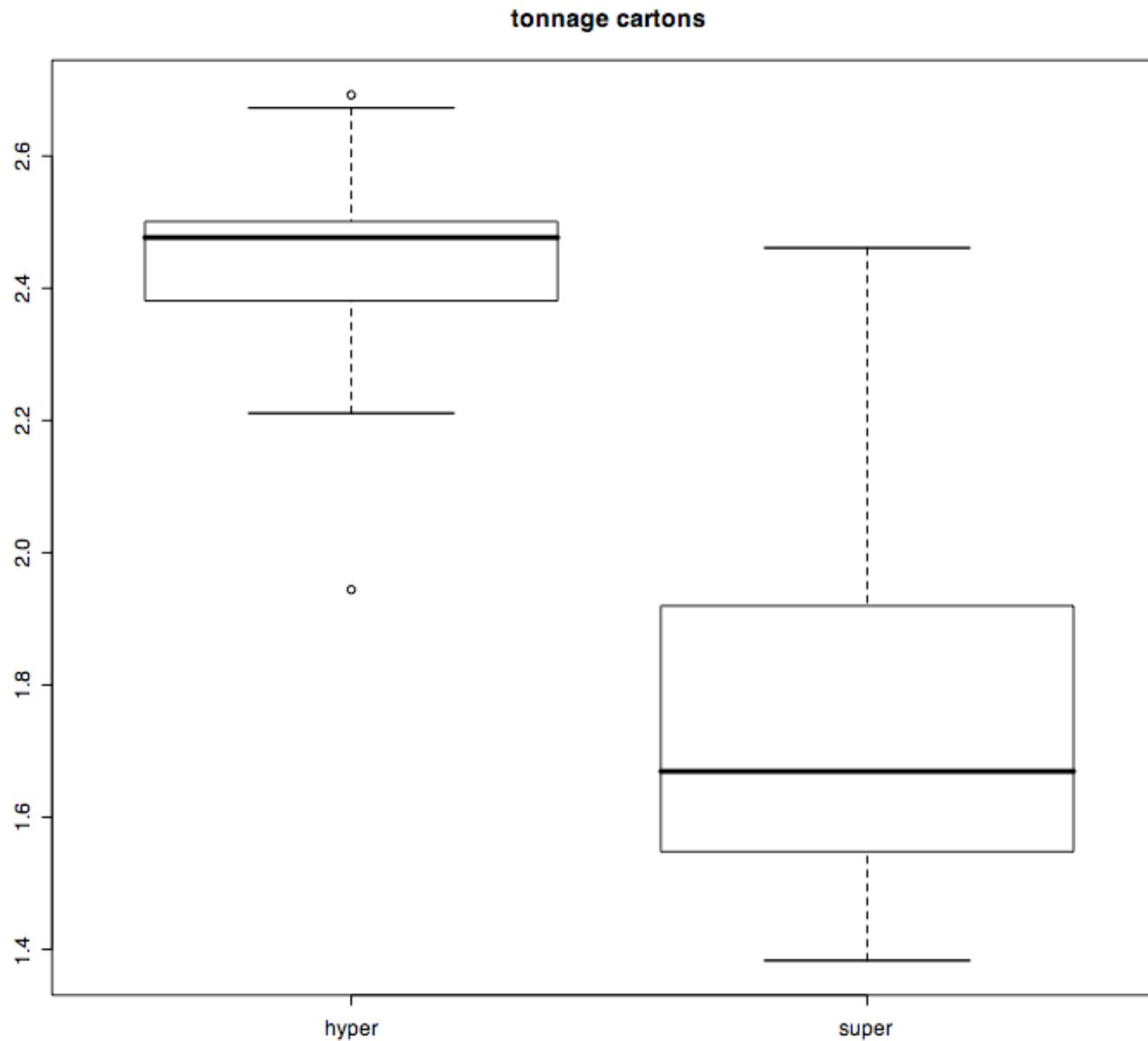
# boxplot(cartons~categorie) - prestataire B



# pairs(cartons~surface+CA) - prestataire B- log10



# boxplot(log10(cartons)~categorie) - prestataire B

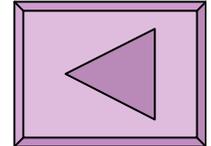


# Modèle pour les log - prestataire B

```
lm(formula = log10(cartons) ~ log10(CA) + log10(surface) +
    categorie,
    data = dechetsB)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.254735	-0.049060	-0.001419	0.041061	0.417644



Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-2.3793	0.6175	-3.853	0.000549	***
log10(CA)	0.9259	0.2139	4.330	0.000145	***
log10(surface)	0.1416	0.2814	0.503	0.618407	
categoriesuper	0.0725	0.1121	0.647	0.522447	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1294 on 31 degrees of freedom

Multiple R-Squared: 0.9052, Adjusted R-squared: 0.896

F-statistic: 98.68 on 3 and 31 DF, p-value: 5.992e-16

# résidus de ce modèle - prestataire B

